Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6346–6353
Marseille, 11–16 May 2020

Parallel Corpus for Japanese Spoken-to-Written Style Conversion

Mana Ihori, Akihiko Takashima, Ryo Masumura
NTT Media Intelligence Laboratories, NTT Corporation
1-1 Hikarinooka, Yokosuka-Shi, Kanagawa 239-0847, Japan
{mana.ihori.kx, akihiko.takashima.dg, ryou.masumura.ba}@hco.ntt.co.jp

Abstract

With the increase of automatic speech recognition (ASR) applications, spoken-to-written style conversion that transforms spoken-style text into written-style text is becoming an important technology to increase the readability of ASR transcriptions. To establish such conversion technology, a parallel corpus of spoken-style text and written-style text is beneficial because it can be utilized for building end-to-end neural sequence transformation models. Spoken-to-written style conversion involves multiple conversion problems including punctuation restoration, disfluency detection, and simplification. However, most existing corpora tend to be made for just one of these conversion problems. In addition, in Japanese, we have to consider not only general spoken-to-written style conversion problems but also Japanese-specific ones, such as language style unification (e.g., polite, frank, and direct styles) and omitted postpositional particle expressions restoration. Therefore, we created a new Japanese parallel corpus of spoken-style text and written-style text that can simultaneously handle general problems and Japanese-specific ones. To make this corpus, we prepared four types of spoken-style text and utilized a crowdsourcing service for manually converting them into written-style text. This paper describes the building setup of this corpus and reports the baseline results of spoken-to-written style conversion using the latest neural sequence transformation models.

Keywords: Spoken-to-written style conversion, Japanese text style conversion, crowdsourcing, ASR transcription

1. Introduction

It has become increasingly important to precisely understand spoken language because various automatic speech recognition (ASR) applications such as artificial intelligence speakers (Li et al., 2017; Purington et al., 2017) and automatic dictation systems (Shang et al., 2018; Li et al., 2019) have been growing recently. Spoken languages are typically transcribed as spoken-style text that includes disfluencies and redundant expressions because ASR systems convert speech into text in a literal manner. However, it is difficult for humans to read such spoken-style text because they are more familiar with reading written-style text that does not include these expressions. In addition, spoken-style text has an adverse effect on subsequent processing (e.g., machine translation, summarization), because these technologies are often developed to handle written-style text. Therefore, spoken-style text needs to be converted into written-style text.

Spoken-to-written style conversion is considered as monolingual translation (Wubben et al., 2010) regarded as sequence to sequence mapping from text to text. So far, various methods such as noisy channel models and hidden Markov models (Johnson and Charniak, 2004; Ferguson et al., 2015; Matusov et al., 2006) have been introduced to handle these kinds of monolingual translation problems. In recent studies, neural sequence transformation models (Sutskever et al., 2014) have been utilized for the monolingual translation problems and demonstrated superior performance (See et al., 2017). However, such models require a large parallel corpus of input text and output text for learning because they directly model the relationship between input text and output text in an end-to-end manner. Thus, to achieve spoken-to-written style conversion using neural sequence transformation models, it is important to prepare a parallel corpus that has a large amount of both spoken-style text and written-style text.

Spoken-to-written style conversion involves multiple conversion problems including punctuation restoration, disfluency detection and simplification. However, conventional studies have handled these conversion problems independently because most existing parallel corpora were made for only one of these conversion problems (Cho et al., 2016; Wang et al., 2016; Pusateri et al., 2017; Tilk and Alumäe, 2016; Pahuja et al., 2017). For example, in disfluency detection problems, the switchboard corpus has mainly been used, in which contains the beginning and ending positions of disfluencies such as fillers and repetitions (Godfrey et al., 1992). When we introduce neural sequence transformation models, a parallel corpus of text including disfluencies and text that eliminates them can be utilized for learning. In addition, in punctuation restoration problems, the IWSLT dataset has mainly been used, in which both punctuated and unpunctuated transcriptions are included (Ueffing et al., 2013). In actuality, the switchboard corpus cannot be utilized for the punctuation restoration problem, and the IWSLT dataset cannot be utilized for disfluency detection problems. In order to handle multiple conversion problems simultaneously, we need to prepare a parallel corpus of spoken-style text and written-style text that include all the conversion problems.

Furthermore, in Japanese, we have to consider not only general spoken-to-written style conversion problems but also Japanese-specific ones. For example, there are several kinds of end-of-sentence expressions in Japanese (e.g., "*desu*" and "*masu*" for polite-style language, "*da*" and "*dearu*" for direct-style language, and "*dayo*" and "*yone*" for frank-style language), but these end-of-sentence expressions should be unified into only one style in written-style text. In addition, postpositional particle expressions are often omitted in spoken-style text. However, they should not be omitted in written-style text to convey the context correctly. Therefore, we need to take into these Japanese fea-

Table 1: Rules of Japanese spoken-to-written style conversion.

Japanese-specific rules		
(1)	Please edit text style in source text to polite-style language such as “*desu*” and “*masu*”.	
	E.g.)	〜みたい → 〜のよう (*# like as*), こっち → こちら (*# this*), とか → など (*# such as*), だよね → ですよね, 〜だったっけ → 〜でしたでしょうか, だったかも → だったかもしれません (polite-style language)
(2)	Please restore postpositional particle expressions (e.g., 〜が, 〜は, 〜に, 〜を), if these are lacking.	
(3)	Please replace English, number, and hiragana notation that are difficult to read with more familiar notations.	
	E.g.)	Kanji notation is better than hiragana notation, alphanumerical notation is better than kanji numerals.
General rules		
(4)	Please restore punctuation and correct restoration errors.	
	E.g.)	Please restore comma (“、”) to appropriate points such as just after conjunctive expressions or when hiragana or kanji notation appears continuously. Please restore period (“。”) to end-of-sentence points.
(5)	Please remove expressions that are not needed like fillers and repetitions.	
	E.g.)	Expressions that are not needed such as ちょっと, あと, はい, あのですね, うーんと, えー.
(6)	Please remove redundant expressions or partition text to read more easily.	
	E.g.)	このデザート、甘いといえば甘いよね → このデザート、甘いですよね。 (*# This dessert is sweet, right?*) 安いと思って、お手軽だと思って買った → 安いと思いました。また、お手軽だと思いましたので買いました。 (*# I thought it was cheap. Moreover, I bought it because it was easy to buy.*)
(7)	Please correct error expressions that are wrongly transcribed considering the context of text.	
	E.g.)	政治のよんさんが足りない → 政治の予算が足りない。 (*# The budget of politics is lacking.*)

tures consideration to make a corpus for Japanese spoken-to-written style conversion.

In this paper, we present a new parallel corpus for Japanese spoken-to-written style conversion that can simultaneously handle general spoken-to-written style conversion problems and Japanese-specific ones. At present, the Corpus of Spontaneous Japanese (Maekawa et al., 2000) has mainly been used for Japanese spoken-to-written style conversion (Tanaka et al., 2019), but fillers, repetitions, and pauses at regular intervals are only annotated to the corpus, and Japanese-specific problems have not been considered at all. To the best of our knowledge, our work is the first attempt to construct a Japanese-specific corpus for spoken-to-written style conversion that considers various conversion problems simultaneously. This paper details how we constructed our corpus, and reports the baseline results of spoken-to-written style conversion using the latest neural sequence transformation models (Luong et al., 2015; See et al., 2017).

Our contributions are three-fold: (1) we designed rules for Japanese spoken-to-written style conversion; (2) we created a parallel corpus for four types of Japanese spoken-style text by utilizing a crowdsourcing service; (3) we investigated the baseline performance of the latest neural sequence transformation models with this created corpus.

2. Related Work

A parallel corpus of spoken-style text and written-style text is beneficial not only for spoken-to-written style conversion but also for written-to-spoken style conversion. Written-to-spoken style conversion can be utilized for language modeling in spontaneous ASR tasks. For example, when using ASR for academic lectures, we can utilize a large amount of written-style text (e.g., proceedings of academic conferences, academic textbooks) for constructing spoken-style language models by converting these text into spoken-style text (e.g., inserting fillers according to rules, utilizing statistical sequence translation) (Hori et al., 2003; Schramm et al., 2003; Akita and Kawahara, 2009; Masumura et al., 2011). Thus, we expect that our corpus can be utilized for building written-to-spoken style conversion based on neural sequence transformation models, and thereby improve the performance in spontaneous ASR tasks.

3. Rules for Japanese Spoken-to-Written Style Conversion

This section details the rules for Japanese spoken-to-written style conversion to construct a Japanese parallel corpus from spoken-style text on a unified basis. These rules are utilized for asking Japanese workers to make written-style text from spoken-style text. First, we defined three rules that focus on Japanese-specific problems. Next, we defined four rules that have been individually utilized in general spoken-to-written style conversion problems. Table 1 summarizes all rules, and Table 2 shows examples of spoken-to-written style conversion by using all rules. In our spoken-style text, only pauses at regular intervals are annotated “<sp>”, as shown in Table 2. In addition, we instructed workers “*Don’t change the content of source text*” because the aim of our corpus is to make a parallel corpus for spoken-to-written *style conversion*. Note that all rules are applied simultaneously.

Table 2: Examples of Japanese spoken-to-written style conversion.

Example 1	
Spoken-style	はいはい<sp>それはそうですね<sp>めたぼが気になるのですか 私なんかは運動をたくさんしているので<sp>ご飯もそれほど食べていないので<sp> だいえっとする必要ってないですね いわゆるメタボとは無縁ちゃ無縁ですが<sp>糖尿病にはきをつけてます
Written-style	それはそうですね。メタボが気になるのですか。 私は、運動をたくさんしていますし、ご飯もそれほど食べません。 よって、ダイエットする必要はないですね。 メタボとは無縁ですが、糖尿病には気を付けてます。
Cf. Translation	That's true. Are you worried about metabolic syndrome? I do a lot of exercise and do not eat so much. So, I do not need to diet. I am not worried about metabolic syndrome, but I am careful about diabetes.

Example 2	
Spoken-style	えっと<sp>もしもし<sp>再配達お願いしていますが<sp>今日は16時までるすにしてます ちょっとですね<sp>申し訳ないのですけれども<sp>16時以降に変更くださいませんか あ<sp>18時以降になっちゃうと出かけますのでそれまでに再配達ができなきゃ<sp> お手数ですが<sp>明日に変更してください こっちの都合で<sp>すいません
Written-style	もしもし、再配達をお願いしていますが、今日は16時まで留守にしてます。 申し訳ないのですが、16時以降に変更させてください。 18時以降になると出かけますので、それまでに再配達ができない場合は、お手数ですが、明日に変更してください。 こちらの都合で、すみません。
Cf. Translation	Hello. I would like to arrange for redelivery, but I am away until 16:00 today. I'm sorry, but I would like to change the delivery time to after 16:00. I will go out after 18:00, so if you can't deliver by 18:00, please deliver tomorrow. Sorry for the inconvenience.

3.1. Japanese-specific rules

(1) Language style unification In Japanese, there are several kinds of end-of-sentence expressions such as "*desu*" and "*masu*" for polite-style language, "*da*" and "*de-aru*" for direct-style language, and "*dayo*" and "*yone*" for frank-style language. It is necessary to be unified into only one end-of-sentence expression in written-style text because if multiple end-of-sentence expressions were used, readers would be confused. Here, direct-style language is used in written-style text only, and frank-style language is used in spoken-style text only. On the other hand, polite-style language is used in both written-style text and spoken-style text. When we convert spoken-style text into written-style text, polite-style language is the most suitable because our spoken-style text is transcriptions of spoken utterances. Therefore, we utilize polite-style language for the written-style text.

Moreover, in Japanese, there are several expressions used in spoken-style text only and used in written-style text only. For example, conjunctive expressions in spoken-style text such as "*demo*" and "*dakara*" should be converted into expressions in written-style text such as "*shikashi*" and "*shitagatte*". It is difficult to complete such conversions accurately because they rely on personal knowledge and experience.

Considering the above, we defined the rule *Please edit text style in spoken-style text to polite-style language such as "desu" and "masu"*, in order to reassure workers who are not used to writing written-style text. In addition, we provided some easy examples, as shown in Table 1.

(2) Postpositional particle expressions restoration In Japanese spoken-style text, postpositional particle expressions are often omitted; however, they should not be omitted in written-style text because we cannot capture the relationships between nouns and verbs or adjectives from the text without them. Therefore, we defined the rule *Please restore postpositional particle expressions such as "ga", "ha", "ni", and "wo", if these are lacking.*

(3) Notation correction In Japanese, there are kanji, hiragana, and katakana notation. Most Japanese are more familiar with kanji notation, which is easier to read than the other notations. On the other hand, notation rules are often not defined when manually or automatically transcribing spoken utterances into spoken-style text, and thus they are often difficult to read. As for numerical characters, alphanumerical characters are easier to read than kanji numerals. For example, in our spoken-style text, hiragana notation has not been converted into kanji notation, katakana notation has been changed to hiragana notation, and alphanumerical characters have been converted into kanji numerals. Therefore, we define the rule *Please replace English, number, and hiragana notations that are difficult to*

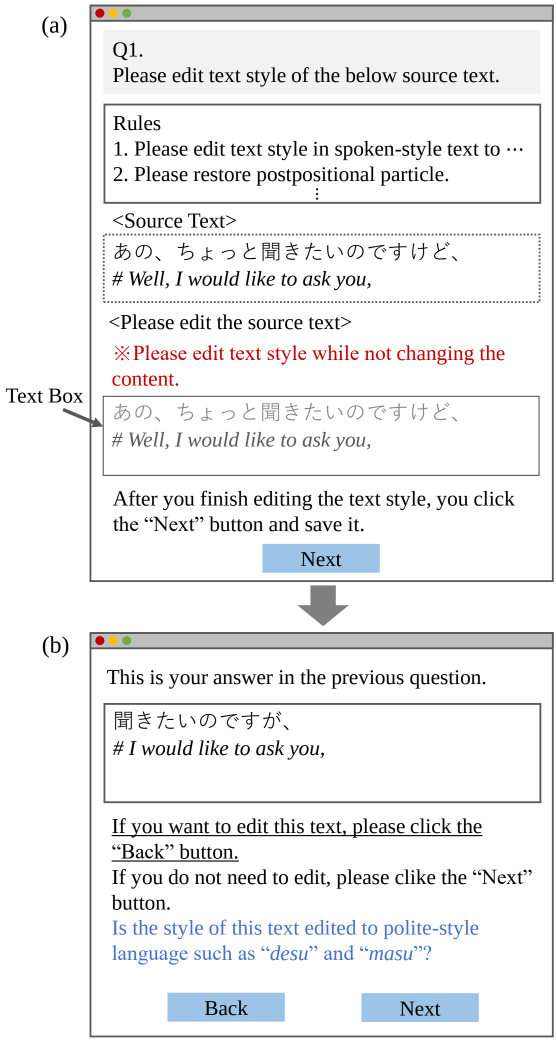

Figure 1: Web page of crowdsourcing platform.

read with more familiar notations, e.g., kanji notation is better than hiragana notation, alphanumerical notation is better than kanji numerals.

3.2. General rules

(4) Punctuation restoration Our spoken-style text has the annotation "<sp>" to indicate pauses at regular intervals; however, these pauses do not always correspond to punctuation marks. In addition, Japanese written-style text requires two kinds of punctuation marks ("、" and "。"). Therefore, we defined the rule *Please restore punctuation in source text (e.g., when hiragana or kanji notation appears continuously, or just after conjunctive expressions), and correct restoration errors.*

(5) Disfluency detection It is difficult to read text that has filler and repetition expressions. Therefore, we defined the rule *Please remove expressions that are not needed like fillers and repetitions.*

(6) Simplification In spoken-style text, redundant expressions are included because converting speech into text in a literal manner. In addition, in spoken-style text, one utterance is often long because humans speak as they think. Therefore, we defined the rule *Please remove redundant expressions or partition text to read more easily.*

(7) Error correction Our spoken-style text has some errors such as verbal slip-up because converting speech into text in a literal manner. Therefore, we defined the rule *Please correct error expressions that are wrongly transcribed considering the context of text.*

4. Corpus Specification

In this section, we describe the corpus specification. To build our corpus, we first prepared four types of spoken-style text, which are transcriptions of Japanese spoken utterances. Next, we hired Japanese workers through a crowdsourcing service and asked them to convert this spoken-style text into written-style text. Finally, we constructed the parallel corpus of spoken-style text and written-style text by filtering noisy data.

4.1. Source spoken-style text

For building the parallel corpus, we collected the following four types of spoken-style text, which are manual transcriptions of spoken utterances.

- Call center dialogue: Simulated call center dialogue datasets between one operator and one customer. We used spoken-style text of both operator and customer. We prepared 3,965 texts in this domain.
- Four-party daily chat: Free daily chat where four people talk about an arbitrary topic such as their hobbies, travel, etc. We prepared 3,962 texts in this domain.
- Two-party daily discussion: Free daily discussion where two people talk about an arbitrary topic such as their life event, their hobbies, etc. We prepared 4,501 texts in this domain.
- Voicemail: Personal voicemail datasets where people leave a message when a phone call cannot be connected. We prepared 12,567 texts in this domain.

We only utilized spoken-style text with more than 20 Japanese characters. The total number of spoken-style text items was 24,995.

4.2. Making written-style text using crowdsourcing

We utilized a crowdsourcing service to hire Japanese workers and convert spoken-style text into written-style text. This service used a Web based questionnaire format, as shown in Figure 1. Here, we displayed the spoken-text in which "<sp>" is replaced with comma "、" to read more easily. We asked the workers to make written-style text by editing source text following the pre-defined conversion rules. Therefore, as shown in (a) on Figure 1, our Web page displayed the conversion rules, source spoken-style text, and a text box to enter the written-style text. Note that the source spoken-style text was initially placed inside the text box to promote editing. After the workers finished editing, we showed them a confirmation page ((b) on Figure 1) to have them confirm their answers. We employed 9,002 Japanese men and women aged 15-88 years

Table 3: Examples of our corpus in call center dialogue domain.

Example 1	
Spoken-style	あすみませんえーとあの解約したいと思って連絡してるんですけどもこちらの番号でよろしいですか
Written-style 1	すみません。解約したいと思って連絡させていただいているのですが。こちらの番号でよろしいでしょうか。
Written-style 2	あの、すみません。解約したいと思って連絡してるんですけど、こちらの番号でよろしいですか。
Written-style 3	すみません。解約したいと思って連絡させていただいたのですが、こちらの番号で間違いございませんでしょうか。
Written-style 4	すみません。解約したいと思って連絡してるのですが、こちらの番号でよろしいですか。
Cf. Translation	Excuse me. I'm calling to cancel the contract. Is this the right phone number?

Example 2	
Spoken-style	はい電話番号が一二三の四五六の七八九〇
Written-style 1	電話番号は、123-456-7890です。
Written-style 2	はい。電話番号が、123-456-7890ですね。
Cf. Translation	OK. My phone number is 123-456-7890.

Table 4: Details of data structures of our corpus.

Domain	Dataset	Number of text spoken	Number of text written
Call center dialogue	Train	2,914	8,169
	Valid	584	584
	Test	465	1,475
Four-party daily chat	Train	2,450	5,328
	Valid	381	381
	Test	361	996
Two-party daily discussion	Train	3,120	8,123
	Valid	581	581
	Test	387	1,150
Voicemail	Train	6,787	15,129
	Valid	1,081	1,081
	Test	1,051	2,794
All domains	Train	15,271	36,749
	Valid	2,627	2,627
	Test	2,264	6,415

for this task. Three or more workers were assigned to one spoken-style text to ensure accurate written-style conversion, as some workers may not edit at all in crowdsourcing. Consequently, each worker was asked to make 10 written-style texts.

Table 3 shows examples of the parallel corpus made by the crowdsourced workers. Note that we manually excluded text (e.g., text that had not been edited at all, text with remaining representative fillers, text without polite-style language), as the quality of data varied depending on personal knowledge and experience. The collected data were divided into a training (Train) set, a validation (Valid) set, and a test set. Table 4 details the corpus with respect to the four source text domains. Note that the total number of spoken-style text and written-style text are different, because a spoken-style text has one or more written-style text.

5. Baselines Evaluation

In this section, we present the baseline results of spoken-to-written style conversion with the created corpus.

Setup We constructed two kinds of networks: an attention-based encoder-decoder network (Luong et al., 2015), a pointer-generator network (See et al., 2017). It is reported that pointer-generator networks can yield a strong performance in monolingual translation tasks because they possess a copy mechanism that appropriately copies tokens from source text to help generate infrequent tokens (Zhang et al., 2018). Here, we trained each network with each domain data and all domain data, as our corpus has four domains. We utilized the following configurations for these networks. In the encoder, a 2-layer bidirectional long short-term memory recurrent neural network (LSTM-RNN) with 512 units was introduced. In the decoder, a unidirectional LSTM-RNN with 512 units was introduced. We used an additive attention mechanism (Bahdanau et al., 2015). We set the output unit size (which corresponds to the amount of characters that appear more than ten times in all training set) to 1,763. To train these networks, we used mini-batch stochastic gradient descent with gradient norm clipping set to 1.0. In each LSTM-RNN, we used dropout and set its rate to 0.2. All trainable parameters were randomly initialized. For the mini-batch training, we truncated each text to 200 characters. The mini-batch size was set to 64. For the decoding, we used a beam search algorithm with the beam size set to four.

Evaluation metrics We calculated automatic evaluation scores for three metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-N and ROUGE-L (RL) (Lin and Hovy, 2003; Lin and Och, 2004). Then, we calculated BLEU-1 (B1), BLEU-2 (B2), and BLEU-3 (B3) with BLEU, and calculated ROUGE-1 (R1),

Table 5: Results of spoken-to-written style conversion by using our corpus.

Domain		B1	B2	B3	R1	R2	RL	METEOR
Call center dialogue	a)	0.736	0.654	0.591	0.717	0.528	0.693	0.737
	b)	0.761	0.693	0.637	0.761	0.594	0.724	0.781
	c)	0.808	0.750	0.699	0.799	0.649	0.770	0.863
	d)	0.813	0.756	0.705	0.803	0.654	0.775	0.867
	e)	**0.826**	**0.772**	**0.723**	**0.811**	**0.666**	**0.784**	**0.886**
Four-party daily chat	a)	0.727	0.654	0.595	0.713	0.542	0.674	0.765
	b)	0.448	0.291	0.195	0.479	0.190	0.339	0.330
	c)	0.757	0.688	0.631	0.755	0.586	0.694	0.820
	d)	0.753	0.680	0.568	0.750	0.580	0.689	0.810
	e)	**0.766**	**0.701**	**0.648**	**0.763**	**0.601**	**0.706**	**0.839**
Two-party daily discussion	a)	0.758	0.685	0.629	0.734	0.560	0.695	0.763
	b)	0.689	0.614	0.557	0.713	0.534	0.650	0.734
	c)	0.765	0.697	0.598	0.761	0.594	0.707	0.818
	d)	0.780	0.712	0.656	0.771	0.606	0.718	0.828
	e)	**0.788**	**0.724**	**0.672**	**0.776**	**0.615**	**0.726**	**0.843**
Voicemail	a)	0.755	0.686	0.629	0.759	0.601	0.731	0.756
	b)	0.811	0.755	0.710	0.820	0.689	0.771	0.845
	c)	0.836	0.789	0.751	**0.842**	**0.726**	0.800	0.882
	d)	0.840	0.789	0.748	**0.842**	0.720	0.799	0.879
	e)	**0.841**	**0.792**	**0.752**	**0.842**	0.723	**0.801**	**0.884**
All domains	a)	0.747	0.674	0.616	0.739	0.570	0.708	0.755
	d)	0.812	0.755	0.707	0.808	0.667	0.764	0.863
	e)	**0.816**	**0.762**	**0.715**	**0.812**	**0.673**	**0.770**	**0.870**

a) Results without introducing any spoken-to-written style conversion networks
b) Attention-based encoder-decoder network trained with only target domain data
c) Pointer-generator network trained with only target domain data
d) Attention-based encoder-decoder network trained with all domain data
e) Pointer-generator network trained with all domain data

ROUGE-2 (R2) with ROUGE-N. Here, we have more than three workers who edit for one source text; in other words, a spoken-style text has more than one written-style text as the correct answer. Thus, we calculate these evaluation scores for all written-style texts.

Results Table 5 shows the results of spoken-to-written style conversion in Japanese. We also evaluated results without introducing any spoken-to-written style conversion networks. As shown, the pointer-generator network trained with all domain data had the best performance. Table 6 shows examples of text that was generated by the pointer-generator network trained with all domain data. These results indicates that it is possible to convert spoken-style text into written-style text by considering both Japanese-specific and general spoken-to-written style conversion problems simultaneously. In addition, Table 5 shows that, in the domains of four-party daily chat and two-party daily discussion, the evaluation results of the attention-based encoder-decoder network trained with only each domain data were lower than the results without introducing any spoken-to-written style conversion networks. We assume that this result was affected by the data quality in each domain. In fact, it is difficult for the network to learn in four-party daily chat because this domain includes a little less text than the other domains, as shown in Table 4. Moreover, while the two-party daily discussion and the call center dialogue have a similar amount of text, the call center dialogue domain has an average of 31 characters per text, which is much lower than the two-party daily discussion domain, whose average is 57 characters per text in each Train. Thus, it is difficult to learn in two-party daily discussion domain because the sequence of this domain is longer than the other domain text. However, we conclude that our corpus can be used to learn neural sequence translational models because the evaluation score of the pointer-generator network trained with all domain data was the highest.

6. Conclusion

In this paper, we have presented a new parallel corpus for Japanese spoken-to-written style conversion featuring four domains. The main difference between our corpus and existing corpora is that ours can simultaneously handle multiple spoken-to-written style conversion problems such as punctuation restoration, disfluency detection and simplification while existing corpora handled only one of these conversion problems. In addition, we dealt with Japanese-specific spoken-to-written style conversion problems such as language style unification (e.g., polite, frank, and direct styles) and omitted postpositional particle expressions restoration. In order to construct our corpus while considering various conversion problems, we de-

Table 6: Examples of automatically generated written-style text.

Call center dialogue	
Input	えーあのもうあんまり使わなくなったということなんですねあの先日からちょっとお習い事を始めまして
Output	あんまり使わなくなったということなんですね。先日からお習い事を始めました。
Cf.Translation	You do not use it too much, don't you? I began a lesson few days ago.
Four-party daily chat	
Input	プロ野球じ野球ほとんどでもルールわかんないですよおふほとんどわかんないですよ高校野球は見ますけど
Output	プロ野球は、ほとんどルールがわかんないんですよ。ほとんどわからないんですよ。高校野球は見ますけど。
Cf.Translation	I do not hardly know the rule of professional baseball; however, I watch high school baseball.
Two-party daily discussion	
Input	でなんかそう言うのをあのあのソフトクリームじゃなくてコンビニとかそう言うとこでも買えたらいいなと個人的にも思っていてでうん
Output	そう言うのをソフトクリームじゃなくて、コンビニとかそう言うとこでも買えたらいいなと個人的にも思っていました。
Cf.Translation	I think that it is convenient if we can buy it at convenience stores.
Voicemail	
Input	あもしもしえーと午前中の打ち合わせ終わりましたので連絡ですえーとー打ち合わせ終わりまして秋葉原の駅まで来ましたえーこれからえー昼食を取りつつえ次の打ち合わせ高田馬場に向かいますえーその連絡でしたえーまた後ほど連絡いたしますよろしくお願いいたします
Output	もしもし、午前中の打ち合わせ終わりましたので、連絡です。打ち合わせが終わりまして、秋葉原の駅まで来ました。これから昼食を取り、次の打ち合わせ、高田馬場に向かいます。また後ほど連絡いたします。よろしくお願いいたします。
Cf.Translation	Hello. I finished the meeting in the morning. And I came to the Akihabara station. I will have lunch, and go to meeting, and next, go to the Takadanobaba station. I would like to call you later. I really appreciate it.

fined both general rules and Japanese-specific rules. To check the quality of our corpus, we evaluated the performance of spoken-to-written style conversion based on the latest neural sequence transformation models. Experimental results showed that pointer-generator networks, which have been used in monolingual machine translation tasks, yield a superior performance, and our trained models can carry out spoken-to-written style conversion while considering both Japanese-specific and multiple general conversion problems. In future work, we will evaluate the performance when applying neural sequence transformation models trained from our corpus to ASR transcriptions. We will also utilize our corpus for building written-to-spoken style conversion models to construct effective language models for ASR.

7. Bibliographical References

Akita, Y. and Kawahara, T. (2009). Statistical transformation of language and pronunciation models for spontaneous speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, pages 1539–1549.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *Proc. International Conference on Learning Representations (ICLR)*, pages 1–15.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.

Cho, E., Niehues, J., Ha, T.-L., and Waibel, A. (2016). Multilingual disfluency removal using NMT. In *Proc. International Workshop on Spoken Language Translation (IWSLT)*.

Ferguson, J., Durrett, G., and Klein, D. (2015). Disfluency detection with a semi-Markov model and prosodic features. In *Proc. the Human Language Technologies Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 257–262.

Hori, T., Willett, D., and Minami, Y. (2003). Language model adaptation using wfst-based speaking-style translation. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages I–I.

Johnson, M. and Charniak, E. (2004). A TAG-based noisy-channel model of speech repairs. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 33–39.

Li, B., Sainath, T., Narayanan, A., Caroselli, J., Bacchiani, M., Misra, A., Shafran, I., Sak, H., Pundak, G., Chin, K., Sim, K. C., Weiss, R. J., Wilson, K., Variani, E., Kim, C., Siohan, O., Weintraub, M., McDermott, E., Rose, R., and Shannon, M. (2017). Acoustic modeling for google home. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Li, M., Zhang, L., Ji, H., and Radke, R. J. (2019). Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2190–2196.

Lin, C.-Y. and Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proc. the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 150–157.

Lin, C.-Y. and Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proc. Annual Meeting on Association for Computational Linguistics (ACL)*, page 605.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421.

Masumura, R., Hahm, S., and Ito, A. (2011). Training a language model using webdata for large vocabulary japanese spontaneous speech recognition. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Matusov, E., Mauser, A., and Ney, H. (2006). Automatic sentence segmentation and punctuation prediction for spoken language translation. In *Proc. International Workshop on Spoken Language Translation (IWSLT)*.

Pahuja, V., Laha, A., Mirkin, S., Raykar, V., Kotlerman, L., and Lev, G. (2017). Joint learning of correlated sequence labelling tasks using bidirectional recurrent neural networks. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 548–552.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proc. Annual Meeting on Association for Computational Linguistics (ACL)*, pages 311–318.

Purington, A., Taft, J., Sannon, S., Bazarova, N. N., and Taylor, S. (2017). "alexa is my new bff": Social roles, user satisfaction, and personification of the amazon echo. In *Proc. Conference Extended Abstracts on Human Factors in Computing Systems (CHI)*, pages 2853–2859.

Pusateri, E., Ambati, B. R., Brooks, E., Platek, O., McAllaster, D., and Nagesha, V. (2017). A mostly data-driven approach to inverse text normalization. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 2784–2788.

Schramm, H., Aubert, X. L., Meyer, C., and Peters, J. (2003). Filled-pause modeling for medical transcriptions. In *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. In *Proc. Annual Meeting of the Association for Computational Linguistic (ACL)*, pages 1073–1083.

Shang, G., Ding, W., Zhang, Z., Tixier, A. J.-P., Meladianos, P., Vazirgiannis, M., and Lorré, J.-P. (2018). Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proc. International Conference on Neural Information Processing Systems (NIPS)*, pages 3104–3112.

Tanaka, T., Masumura, R., Moriya, T., Oba, T., and Aono, Y. (2019). Disfluency detection based on speech-aware token-by-token sequence labeling with blstm-crfs and attention mechanisms. In *Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1009–1013.

Tilk, O. and Alumäe, T. (2016). Bidirectional recurrent neural network with attention mechanism for punctuation restoration. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3047–3051.

Wang, S., Che, W., and Liu, T. (2016). A neural attention model for disfluency detection. In *Proc. International Conference on Computational Linguistics (COLING)*, pages 278–287.

Wubben, S., Van Den Bosch, A., and Krahmer, E. (2010). Paraphrase generation as monolingual translation: Data and evaluation. In *Proc. International Natural Language Generation Conference (INLG)*, pages 203–207.

Zhang, Z., Li, W., and Sun, X. (2018). Automatic transferring between ancient chinese and contemporary chinese. In *Proc. Natural Language Processing and Chinese Computing (NLPCC)*, pages 157–167.

8. Language Resource References

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 517–520.

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous speech corpus of japanese. In *Proc. International Conference on Language Resources and Evaluation (LREC)*, pages 947–9520.

Ueffing, N., Bisani, M., and Vozila, P. (2013). Improved models for automatic punctuation prediction for spoken and written text. In *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pages 3097–3101.